How Do Professors Format Exams?
An Analysis of Question Variety at Scale

Paul Laskowski, Sergey Karayev, and Marti Hearst
Learning at Scale 2018
WHO

Paul Laskowski
UC Berkeley
Professor at School of Information

Sergey Karayev
Gradescope, Inc.
Co-founder, Director of Research

Marti Hearst
UC Berkeley
Professor at School of Information & CS
• We analyzed questions on nearly 1,800 paper exams graded in real STEM courses

• We annotated the type of each question (multiple choice, short/medium/long writing, drawing)

• We found interesting differences by type in position, use across subjects, student performance, and reliability
  • e.g. 5+ binary choice questions are required to match reliability of 1 long writing question
Why

• Unique vantage point on exams “in the wild”
• Near term: get an understanding of what is happening today
• Longer term: improve exam writing for both instructors and students
• The trend: oral open-response -> written open-response -> multiple-choice
• Good assessment is valid and reliable.
  • Ideally, summative assessment is also discriminative and easy to grade without bias.
• Many studies on open-response vs. multiple-choice
  • Well-written MCs can be as good as essays
  • MCs are hard to write, easy to grade. Essays are the opposite.
• Ordering effects are unclear
Gradescope: a Fast, Flexible, and Fair System for Scalable Assessment of Handwritten Work
Arjun Singh, Sergey Karayev, Kevin Gutowski, Pieter Abbeel
Learning at Scale 2017

DATA SOURCE
• No constraints on type of questions
• Consistent scoring due to shared, modifiable rubric
40M+ student answers
  • graded by 6,000+ instructors (mostly STEM)
  • at 600+ schools (mostly US higher-ed)
• Built special annotation interface, hired educators
• Annotated over 50% of 50,000+ questions that had at 100+ student answers at the time
<table>
<thead>
<tr>
<th>RESPONSE AREA</th>
<th>TYPICAL RESPONSE LENGTH</th>
<th>RESPONSE TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Other</td>
<td>N/A</td>
<td>Binary MC</td>
</tr>
<tr>
<td>Bubble or Square MC</td>
<td>Character, word, or words</td>
<td>Multiple MC</td>
</tr>
<tr>
<td>Other kinds of MC</td>
<td>Sentence or a couple of lines</td>
<td>Text</td>
</tr>
<tr>
<td>Small fill-in-the-blank</td>
<td>Paragraph or more</td>
<td>Code</td>
</tr>
<tr>
<td>Line-sized</td>
<td></td>
<td>Math</td>
</tr>
<tr>
<td>Paragraph-sized</td>
<td></td>
<td>Chemical Structure</td>
</tr>
<tr>
<td>Page-sized</td>
<td></td>
<td>Self-boxed</td>
</tr>
<tr>
<td>Matrix or grid</td>
<td></td>
<td>Justification/Extension</td>
</tr>
<tr>
<td>Diagram, plot, or chem structure</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Box for drawing</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Final Dataset

• Almost 23,000 questions
• From almost 1,800 exams
• Corresponding to graded answers of 120,000+ students
QUESTION TYPE SAMPLES

(c) The polynomial $x - 1 \mod 15$ has at most 3 zeros.
   True [Circle] False

(ii) [2 pts] Continuous
   - Self-Driving Car [Circle]
   - Pacman

(a) (4 points) $Q_0 = 0 \quad Q_0 \neq 0 \quad Q_2 = 1 \quad Q_2 \neq 1 \quad Q_3 \neq k$

2) What protocol is used to find the hardware address?
   a) RARP
   b) ICMP
   c) None of the above

3. Complete the following:
   ```c
   typedef struct blo {
       item_t *head;
       pthread_mutex_t blo_lock;
   } blo_t;
   ```

4. What are the two species present at high concentration?
   Nac1, NaNO2

5. (c) Pre-image resistance (first pre-image resistance)
   Given a hash value $h$, it should be computationally infeasible to find $x$ such that $h = h(x)$

(c) (2 pts) What is the distance to default?
   Distance to default is defined as $\ln(\frac{b}{A})$
   Since the default level is 90% of the value from value $\frac{1}{0.9} = \frac{1111}{1} = 1.11111\ldots$
   $\ln(\frac{b}{A}) = 0.105$

6. (c) Short Writing

7. (5 pts) SGD-based training demonstrates its benefits in various applications. What are these benefits?

8. (3 pts) Draw a 7p orbital. Show all nodes and phasing for full credit (6 pts)

9. (4 pts) Use the HA-tag to purify your favorite protein. Draw a diagram showing the purification setup.
(c) The polynomial $x - 1 \mod 15$ has at most 3 zeros.

True  False

(ii) [2 pts] Continuous

- Self-Driving Car
- Pacman

(a) (4 pts) Binary

2) What is a pre-image resistance in a cryptographic function?

(e) Pre-image resistance (first pre-image resistance)
zeros.

(a) (4 points) \( A_0 = 0 \)  \( A_1 = 1 \)  \( A_k = k \)
\( A_0 \neq 0 \)  \( A_1 \neq 1 \)  \( A_k \neq k \)

2) What protocol is used to find the hardware address?
   a) RARP
   b) ICMP
   c) None of the above

b) Multiple Choice
d. What are the two species present at high concentration?

\[ \text{Nal, NaNO}_2 \]

3. Complete the following:

```c
typedef struct blo {
    item_t *head;
    pthread_mutex_t blo_lock;
} blo_t;
```

c) Short Writing
(c) Pre-image resistance (first pre-image resistance)

Given a hash value \( h \), it should be computationally infeasible to find \( x \) such that \( h = h(x) \).

(c) (2 pts) What is the distance to default?

Distance to default is defined as \( \ln \left( \frac{B}{V} \right) \).

Since the debt level is 90% of the value from value \( \frac{1}{0.9} = \frac{V}{B} = 1.1111 \ldots \).

\[ \ln \left( \frac{1}{0.9} \right) = 0.105 \]

(d) Medium Writing
1. (6 points) An unknown compound contains carbon, hydrogen, and oxygen. A 25.0 g sample was burned to produce 30.45 g of carbon dioxide and 14.29 g of water. What is the formula of the unknown compound?

\[ 36.82 \text{ g } \text{C}_2 \times \frac{12.01 \text{ g}}{16.00 \text{ g}} = 24.05 \text{ g } \text{C} \]
\[ 19.97 \text{ g } \text{H}_2 \text{O} \times 2.02 \text{ g } \text{H}_2 \text{O} \times \frac{1 \text{ mol } \text{H}_2 \text{O}}{18.02 \text{ g } \text{H}_2 \text{O}} = 1.10 \text{ mol } \text{H}_2 \text{O} \]
\[ 9.97 \text{ g } \text{C} \times \frac{1 \text{ mol } \text{C}}{12.01 \text{ g } \text{C}} = 0.83 \text{ mol } \text{C} \]
\[ 1.6812 \text{ g } \text{H}_2 \text{O} \times \frac{1 \text{ mol } \text{H}_2 \text{O}}{18.02 \text{ g } \text{H}_2 \text{O}} = 0.092 \text{ mol } \text{H}_2 \text{O} \]
\[ 3.52 \text{ g } \text{O}_2 \times \frac{3 \text{ mol } \text{O}_2}{16 \text{ g } \text{O}_2} = 0.21 \text{ mol } \text{O}_2 \]
\[ \frac{0.1323}{0.2021} = 0.65 \]
\[ 1.645 \times \frac{0.2024}{0.2021} = 8 \text{ H} \]
\[ 0.2024 \]
\[ \text{C}_4 \text{H}_8 \text{O} \]

4. (20 pts) Validation is a key concept in Minsky’s financial instability hypothesis. Write an essay to explain this concept. Pay attention to how validation influences lending standards. Include at least one example of validation from the subprime lending crisis. What are the implications of validation leading up to the Great Recession for future business cycles?

The hypothesis central to the subprime mortgage crisis was the belief that default rates would be low. Because of this belief, many lenders were willing to extend credit to people who were not creditworthy, leading to a decrease in the quality of the mortgage-backed securities. This belief was validated by the low default rates in the early years of the mortgage crisis. However, as the economy weakened, the quality of the credit extended by these lenders diminished, and the market for mortgage-backed securities collapsed. This collapse led to a significant increase in the price of mortgage-backed securities, which in turn led to a decrease in the availability of credit. As a result, the ability of many people to purchase homes and start businesses was severely limited.
3. Draw a 7p orbital. Show all nodes and phasing for full credit (6 pts)

A). (4 pts) Use the HA-tag to purify your favorite protein. **Draw a diagram** showing the purification setup.

f) Drawing
How do scores differ by question type?
### Mean Scores by Question Type

<table>
<thead>
<tr>
<th>Type</th>
<th>Score</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>binary choice</td>
<td>0.78</td>
<td>(0.17)</td>
</tr>
<tr>
<td>short writing</td>
<td>0.74</td>
<td>(0.21)</td>
</tr>
<tr>
<td>other</td>
<td>0.73</td>
<td>(0.20)</td>
</tr>
<tr>
<td>multiple choice</td>
<td>0.70</td>
<td>(0.20)</td>
</tr>
<tr>
<td>drawing</td>
<td>0.70</td>
<td>(0.18)</td>
</tr>
<tr>
<td>medium writing</td>
<td>0.69</td>
<td>(0.21)</td>
</tr>
<tr>
<td>multi-type</td>
<td>0.68</td>
<td>(0.19)</td>
</tr>
<tr>
<td>long writing</td>
<td>0.67</td>
<td>(0.20)</td>
</tr>
</tbody>
</table>
Researchers have reported on the differences in discriminability between True-False (TF), or binary choice questions and MC item types. For instance, as far back as 1941, Cronbach [5] showed experimentally the bias of students towards choosing True in a TF choice. Frisbie [9] found in one experiment that TF questions were significantly less reliable than MC items on the same subject material. But Ebel [7] found in a controlled experiment that teachers could get about the same degree of discriminability with TF questions as with multiple choice questions if they wrote five TF questions for every three MC questions.

For this dataset, Table 2 presents the detailed counts and average score mean and standard deviation for Binary, 1-of-N (exactly one choice must be selected), and K-of-N (several choices may be selected) multiple choice questions. The binary type has the highest mean score, followed by 1-of-N, and then K-of-N.

Over 95% of all multiple choice questions offered eight or fewer choices. The average number of choices for 1-of-N questions was 5.0; the average for K-of-N questions was 5.7. The number of choices was inversely correlated with the average score on the question (correlation strength of 0.18), ranging from 78.6% mean score for two choices to 66.0% mean score for eight choices.

K-of-N multiple choice questions are inherently more difficult due to the expanded answer space, reducing the likelihood that a guess will be correct, but they are also more amenable to assigning partial credit to student answers. We found that 33.0% of answers to K-of-N questions were given partial credit by instructors, in comparison to only 7.7% of answers to 1-of-N questions.
How do instructors sequence questions?
• Mean scores drop with position in exam.
FIRST HALF VS SECOND HALF

• More frequently occur in first half of exam:
  • Binary choice (70%)
  • Multiple choice (56%)
  • Short writing (53%)

• More frequently occur in second half of exam:
  • Long writing (60%)
  • Medium writing (53%)
Mean scores drop with position in exam for all question types.
HOW RELIABLE ARE DIFFERENT QUESTION TYPES?
reliability = \frac{\text{signal}}{\text{signal} + \text{noise}}
**Question Reliability**

Ideal: give each student an infinite number of questions

\[
\text{reliability} = \frac{\text{signal}}{\text{signal + noise}} = \frac{\text{variance in student ability}}{\text{total score variance}}
\]
**Question Reliability**

Ideal: give each student an infinite number of questions

\[
\text{reliability} = \frac{\text{signal}}{\text{signal} + \text{noise}} = \frac{\text{variance in student ability}}{\text{total score variance}}
\]

Estimate with a linear mixed model (Hoyt 1947, Cronbach 1951)

**Observed data**
Our Model

\[ \text{score}_{s,q,a,t} = G_t + K_{s,t} + A_{s,a} + u_{s,q,a,t} \]

- **Global question type average**
- **Student-assignment affinity**
- **Student aptitude for the type**
- **Student-question affinity**

Observed score of student s on question q in assignment a of type t
Our Model

\[ \text{score}_{s,q,a,t} = G_t + K_{s,t} + A_{s,a} + u_{s,q,a,t} \]

- Key assumption is that each effect is independent.
- Many factors not accounted for:
  - Students that work hard inspire others to work hard.
  - Instructors make up for hard questions/assignments with easy ones.
  - Consecutive questions cover similar material.
## Reliability Estimates

<table>
<thead>
<tr>
<th></th>
<th>Reliability</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Binary</strong></td>
<td>0.036</td>
<td>0.002</td>
</tr>
<tr>
<td><strong>Multiple Choice</strong></td>
<td>0.061</td>
<td>0.001</td>
</tr>
<tr>
<td><strong>Short Writing</strong></td>
<td>0.112</td>
<td>0.001</td>
</tr>
<tr>
<td><strong>Medium Writing</strong></td>
<td>0.120</td>
<td>0.001</td>
</tr>
<tr>
<td><strong>Long Writing</strong></td>
<td>0.170</td>
<td>0.001</td>
</tr>
<tr>
<td><strong>Drawing</strong></td>
<td>0.144</td>
<td>0.002</td>
</tr>
<tr>
<td><strong>Multi-type</strong></td>
<td>0.117</td>
<td>0.001</td>
</tr>
</tbody>
</table>
EBEL, R. L. Can teachers write good true-false test items?
To what degree are different students suited for different question types?
$$score_{s,q,a,t} = G_t + K_{s,t} + A_{s,a} + u_{s,q,a,t}$$

- Observed score of student $s$ on question $q$ in assignment $a$ of type $t$
- Global question type average
- Student aptitude for the type
- Student-assignment affinity
- Student-question affinity
<table>
<thead>
<tr>
<th></th>
<th>Binary</th>
<th>MC</th>
<th>Short</th>
<th>$t = \text{Medium}$</th>
<th>Long</th>
<th>Drawing</th>
<th>Multi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$s = \text{Student B}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Student C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Student D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Student E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Question Type Aptitude Correlations

<table>
<thead>
<tr>
<th>Question Type</th>
<th>Binary Choice</th>
<th>Multiple Choice</th>
<th>Short Writing</th>
<th>Medium Writing</th>
<th>Long Writing</th>
<th>Drawing</th>
<th>Multi-Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple Choice</td>
<td>0.753</td>
<td></td>
<td>0.78</td>
<td>0.791</td>
<td>0.762</td>
<td>0.794</td>
<td>0.806</td>
</tr>
<tr>
<td>Short Writing</td>
<td></td>
<td>0.791</td>
<td>0.78</td>
<td>0.794</td>
<td>0.805</td>
<td>0.806</td>
<td>0.805</td>
</tr>
<tr>
<td>Medium Writing</td>
<td></td>
<td></td>
<td>0.794</td>
<td>0.813</td>
<td>0.915</td>
<td>0.913</td>
<td>0.915</td>
</tr>
<tr>
<td>Long Writing</td>
<td></td>
<td></td>
<td></td>
<td>0.813</td>
<td>0.915</td>
<td>0.995</td>
<td>0.915</td>
</tr>
<tr>
<td>Drawing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.884</td>
<td>0.921</td>
<td>0.884</td>
</tr>
<tr>
<td>Multi-Type</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.97</td>
<td>0.95</td>
<td>0.97</td>
</tr>
</tbody>
</table>
Eigenvalue:
6.09
<table>
<thead>
<tr>
<th>Component Value</th>
<th>first</th>
<th>second</th>
<th>third</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eigenvalue:</td>
<td>6.09</td>
<td>0.37</td>
<td>0.27</td>
</tr>
<tr>
<td>binary</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>short</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>medium</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>long</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>drawing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>multi-type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>binary</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>short</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>medium</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>long</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>drawing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>multi-type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>binary</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>short</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>medium</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>long</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>drawing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>multi-type</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
SUMMARY

• Revealed some common patterns in university STEM exams today
  • Start has binary and multi-MC, ending has open response
  • Mean score drops with position in exam
• Found significant differences in question type reliability
• Found that student aptitude for question types is largely correlated
  • But: binary/multiple choice questions are less correlated with others
• More interesting findings in the paper!
THANK YOU!

LET'S COLLABORATE:

SERGEY@GRADESCOPE.COM